Gene 537 (2014) 29–40

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Analysis of changes in transcription start site distribution by a classification approach

Kuo-ching Liang [a], Yutaka Suzuki [c], Yutaro Kumagai [d], Kenta Nakai [a,b,*]

[a] Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[b] Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba-ken 227-8561, Japan
[c] Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba-ken 227-8561, Japan
[d] Laboratory of Host Defense, World Premier International Immunology Frontier Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan

ARTICLE INFO

Article history:
Accepted 16 December 2013
Available online 31 December 2013

Keywords:
Changes of TSS distribution
Statistical approach
Time-course TSS data
Alternative promoters
Mouse dendritic cells

ABSTRACT

Change in transcription start site (TSS) usage is an important mechanism for the control of transcription process, and has a significant effect on the isoforms being transcribed. One of the goals in the study of TSS is the understanding of how and why their usage differs in different tissues or under different conditions. In light of recent efforts in the mapping of transcription start site landscape using high-throughput sequencing approaches, a quantitative and automated method is needed to process all the data that are being produced. In this work we propose a statistical approach that will classify changes in TSS distribution between different samples into several categories of changes that may have biological significance. Genes selected by the classifiers can then be analyzed together with additional supporting data to determine their biological significance. We use a set of time-course TSS data from mouse dendritic cells stimulated with lipopolysaccharide (LPS) to demonstrate the usefulness of our method.

1. Introduction

With recent advances in the understanding of complex mechanisms involved in the regulation of transcription in eukaryotes, our view of gene transcription landscape has changed dramatically. At the completion of Human Genome Project, the number of genes identified (~20,000) was far smaller than what was previously estimated (~50,000 to ~100,000). Subsequent studies have shown that in order to produce the large number of known proteins from the smaller than expected set of genes, a gene will often produce multiple unique isoforms, accomplished through several different mechanisms (Landry et al., 2003). In particular, production of multiple isoforms due to usage of alternative promoters, which was once considered as uncommon, has now been found to be a mechanism involved in the majority of human genes (Davuluri et al., 2008; The ENCODE Project Consortium, 2012). The analysis of alternative promoter has become an important topic in the study of transcriptional machinery, not only to find genes with alternative isoforms, but also to understand the evolutionary history of regulatory and transcriptional mechanism for these genes (Jordan et al., 2003).

The usage of alternative promoters can result from changes in epigenetic modifications such as DNA methylation, histone modifications and chromatin remodeling, or from changes to using different transcription factors that bind to different promoters (Hatchwell and Greally, 2007).

Abbreviations: AP-1, jun proto-oncogene; *BAI3*, brain-specific angiogenesis inhibitor 3; *CADM2*, cell adhesion molecule 2; CAGE, cap analysis of gene expression; cDNA, DNA complementary to RNA; ChIP, chromatin immunoprecipitation; *FGF14*, fibroblast growth factor 14; *GADD45g*, growth arrest and DNA-damage-inducible 45 gamma; GM-CSF, Granulocyte macrophage colony-stimulating factor; *IFIT1*, interferon-induced protein with tetratricopeptide repeats 1; *IFNR*, interferon production regulator; *IKK*, inhibitor of kappa B kinase; *IKZF1*, IKAROS family zinc finger 1; *IL6*, interleukin 6; *IL27*, interleukin 27; *IRAK*, interleukin receptor-associated kinase; *IRF*, Interferon regulator factor; *ISG15*, ubiquitin-like modifier; JAK/STAT, Janus kinase/signal transducers and activators of transcription; *KCNIP4*, K_v channel interacting protein 4; K_v, voltage-gated potassium channel; LPS, lipopolysaccharide; *IRG1*, immunoresponsive gene 1; *LRRTM4*, leucine rich repeat transmembrane neuronal 4; *LSAMP*, limbic system-associated membrane protein; *MAPK*, mitogen-activated protein kinase; *MyD88*, myeloid differentiation primary response gene 88; *NF-κB*, nuclear factor kappa B; *NFKBIZ*, nuclear factor of kappa light polypeptide gene enhancer in B cells inhibitor, zeta; *NRG3*, neuregulin 3; *NRXN1*, neurexin 1; *PCDH9*, protocadherin 9; *RIP1*, receptor (TNFRSF)-interacting serine-threonine kinase 1; *SOCS1*, suppressor of cytokine signaling 1; *STAT5a*, signal transducer and activator of transcription 5A; *TAK1*, *TGF-beta*-activated kinase 1; *TANK*, *TRAF* family member-associated *NF-κB* activator; *TBK1*, *TANK*-binding kinase 1; *TLR4*, toll-like receptor 4; *TNF-α*, Tumor necrosis factor alpha; *TRAF*, *TNF* receptor associated factors; *TRAFD1*, TRAF-type zinc finger domain containing 1; *TRIF*, toll-like receptor adaptor molecule; TSC, transcription start site cluster; TS, tissue specificity score; TSS, transcription start site; *TTR*, transthyretin; *USP18*, ubiquitin specific peptidase 18.

* Corresponding author.
E-mail address: knakai@ims.u-tokyo.ac.jp (K. Nakai).

0378-1119
http://dx.doi.org/10.1016/j.gene.2013.12.038

These two mechanisms allow genes to utilize different promoters through different means: one by blocking access to promoters, forcing transcription factors to find different binding targets, the other by using different transcription factors to bind to different targets. Genes with possible alternative promoter usage under different conditions can be found by analyzing promoter binding or transcription start site data. In this work we focus our efforts on the analysis of transcription start sites.

Many computational methods (Bajic et al., 2002; Down and Hubbard, 2002; Knudsen, 1999; Lu and Luo, 2008; Zhang, 1998) and experimental approaches such as Cap Analysis of Gene Expression (CAGE), massively parallel Paired End Tag (PET)-tagging, and TSS-Seq have been proposed to identify TSS and the corresponding promoters (Birney et al., 2007; Carninci et al., 2005; Suzuki et al., 2001; Tsuchihara et al., 2009). Recent cDNA sequencing projects such as FLJ (Ota et al., 2004) and FANTOM (Okazaki et al., 2002) have revealed that instead of utilizing only a single TSS, a promoter can be associated with a number of TSS that are distributed around its immediate neighborhood. Databases such as DBTSS (Yamashita et al., 2012) have made public up to 418 million TSS tags generated using oligo-capping and TSS-Seq techniques, providing a comprehensive overview of TSS landscapes and allowing for their comparisons in tissues under different conditions. The understanding of how the distribution of TSS changes under different conditions can help to shed further insight into the mechanism for transcribing different isoforms, and possibly their differences in functions.

Researchers have already begun to explore the relationship between TSS and transcription mechanisms. Some take the integrative approach where RNA-Seq and ChIP-Seq data are utilized in the analysis of TSS data (Yamashita et al., 2011). Others have taken the approach to analyze the significance of differences in TSS distributions. In Carninci et al. (2006), distributions of TSS are classified into four groups: 1) Single dominant peak, 2) Broad, 3) Bi- or multi-modal, and 4) Broad with dominant peak; and shapes of TSS distributions are correlated to nucleotide sequences and expression levels in human and mouse. In particular, TSS distributions with a single dominant peak are often associated with promoters with TATA-box motif, whereas broad distributions are typically found in promoter regions that have high CG content or are enriched with CpG islands (Gustincich et al., 2006). Other similar classification systems based on shapes of TSS distributions have also been proposed (Ni et al., 2010). However, while characterizing TSS distribution based on shapes of distributions has revealed some correlation with gene expression, the heuristic-based approach in determining the type of distribution shape may be a limiting factor in the uncovering of more complex relationships. In Yamashita et al. (2011), genes with TSS distribution changes in different tissues are grouped into categories based on the pattern of distribution differences, and functional overrepresentations are identified from gene ontology analysis for each category. These findings highlight the utility of not simply looking at whether differences in tag distributions exist between samples, but also taking one step further in identifying genes with specific kinds of distribution change patterns that are of interest for the given study. Furthermore, with advances in sequencing technology that allow researchers to generate TSS data in an unprecedented quantity and speed, a need has arisen for statistical methods that can automatically compare TSS distributions between different samples to identify such unique patterns.

Currently, there are many well-established methods that can be used to detect differential expression in RNA-Seq analyses. For example, in *edgeR* (Robinson et al., 2010) and *DESeq* (Anders and Huber, 2010), read count of a gene, transcript, or exon is modeled as a negative binomial distribution. In both methods the mean and variance of a negative binomial distribution are modeled as functions of the true relative abundance, due to the often lack of samples to estimate variance separately. Thus, differential expression is detected by testing the null hypothesis that the true relative abundances are the same in different samples. However, such methods pool all the reads into a single read count, and provide no information regarding how the reads are mapped to different parts of the gene/transcript/exon, and whether the distribution of these mappings are different between the samples being compared. In Kawaji et al. (2006), differences in the distribution of CAGE tags for TSS are categorized into positional bias and regional bias. For positional bias, Kruskal–Wallis one-way analysis of variance is used to test the null hypothesis that a gene's TSS distributions in different samples have the same median. For regional bias, a tissue specificity score (TS) is computed for each 21-bp window. High TS indicates that the tissue has a tissue-specific preference for TSS usage in this 21-bp region compared to other tissues. However, the Kruskal–Wallis test does not actually test for equal median or mean, and may give inaccurate results when the distribution have different shapes [Handbook of Biological Statistics]. Furthermore, while TS can locate regional differences in tag distributions, it is unclear how TS from various regions can be combined to give a single score to represent how well the overall distribution change matches the change pattern of interest. In (Zhao et al., 2011), Minimum Difference of Pair Assignments, which is similar to Earth Mover's Distance (Rubner et al., 1998), is proposed to compare the similarity between TSS distributions. However, this is again a global measure of difference between distributions, and does not contain any information on the pattern of the difference between the distributions. In (Balwierz et al., 2009), TSS loci are grouped into TSS clusters (TSC), and the likelihood was derived for two neighboring TSCs under the assumption that they have fixed relative expression. While this approach provides a comparison of the proportionality of adjacent TSCs, its computation may become overly complex when we want to make a gene-level comparison where many TSCs may be involved. Furthermore, in a multi-sample comparison, the approach cannot distinguish in which sample the change in TSC expression has occurred, and in a two-sample comparison, the likelihood function may not be accurately estimated.

In this work, we propose a classifier that can be reconfigured to test for specific patterns of TSS distribution change between tissues. We will use this approach to construct classifiers to identify genes that show differential expression in two different samples while utilizing the same TSS, and genes that exhibit TSS shift between two different samples, which we name Class 1 and Class 2 genes, respectively. The pattern of distribution change of Class 2 genes is of particular interest in our analysis of TSS, due to the possible link to alternative promoter usage, and the unavailability of such information in traditional transcriptome analysis such as microarrays and RNA-Seq. The proposed classifier analyzes TSS distributions in different samples by directly comparing their distributions in high resolution, using only a user-defined window size to merge TSS loci that might be using the same promoter. To test its usefulness, we will apply the proposed classifier to a set of TSS-Seq data for a time-course experiment on mouse dendritic cells to discover genes with possible alternative promoter usage after stimulation. It should be noted here that the classifier proposed in this paper is for single sample experiment only. While in recent years many works have argued that noise found in biological replicates is significant enough to put doubt in findings from single sample experiments as to whether statistical significant findings are due to biological phenomenon or within sample variations, when used with caution, single sample experiments can still be informative in a preliminary manner, providing candidates for more in-depth follow-up studies. In particular, many databases, including DBTSS, which is one of the largest repositories of sequencing data for TSS, contain many single sample experiments, and analyses of these data can still provide valuable knowledge about the mechanism for transcription.

1.1. TLR signaling pathways

An important motivating application for the TSS distribution change classifier is for the understanding of potential changes in TSS usage when a dendritic cell is being stimulated by lipopolysaccharide (LPS). Dendritic cells act as intermediaries between external environment

and mammalian adaptive immunity mechanism by presenting foreign antigens to various types of lymphocytes. Deciphering the complex web of control and interacting relationships in dendritic cells is therefore of critical importance in the understanding of mammalian immune system. An important pathway involved in the activation of innate immune response is the Toll-like receptor 4 (TLR4) signaling pathway. The binding of LPS, which is found in the outer membrane of Gram-negative bacteria, to the extracellular domain of TLR4 activates a chain of signal that eventually leads to the activation of proinflammatory cytokines and Type I interferons (Takeda and Akira, 2004).

After activation by LPS binding, cytoplasmic domain of TLR4 recruits adaptor proteins to assist in the subsequent signal transduction. The two adaptor proteins involved in TLR4 signaling pathway are MyD88 and TRIF. Both MyD88 and TRIF react to the activation of TLR4 by modulating downstream signaling pathways. Experimental results have shown that MyD88 and TRIF each mediate a pathway that is independent from the other (Kawai et al., 2001). In MyD88-dependent pathway, MyD88 recruits and activates additional proteins such as IRAK-4, IRAK-1, TRAF6 and TAK1, which leads to the activation of IκB kinase (IKK) and mitogen-activated protein kinase (MAPK) pathways, and eventually the activation of NF-κB and AP-1 transcription factors, which play roles in the expression of proinflammatory cytokines. On the other hand, in MyD88-independent pathway (or TRIF-dependent pathway), TRIF activates RIP1 and the IKKs, which lead to the expression of NF-κB transcription factors. TRIF, through TRAF3, TANK, TBK1 and IKKi, also activates IRF3. Both NF-κB and IRF3 are important in the induction of Type I interferons (Lu et al., 2008). It has also been shown that certain proinflammatory cytokines, such as Il6 and TNF-α, are dependent on both MyD88 and TRIF mediated pathways (Shen et al., 2008).

To further understand the mechanisms in which the TLR signaling pathway activates innate immune response in dendritic cells after being stimulated by LPS, we would like to find out which genes are activated after an immune response, when do these activated genes reach their peak expressions, how the time-course expression patterns of these genes cluster, and whether these differences and similarities agree with known networks or pathways or could shed light on novel relationships. In particular, we are interested in mapping out how each gene's TSS distribution changes over time after LPS stimulation. By observing the dynamical behavior of a gene's TSS usage before and after LPS stimulation, and by classifying the observed TSS distribution changes according to a pre-specified pattern, we not only can determine whether alternative promoter usage is part of the mechanism in the TLR signaling pathway, but also can determine whether the detected TSS usage changes affect the time-course expressions of the genes involved in the TLR signaling pathway.

2. Materials and methods

2.1. Methods

To compare the TSS tag distributions of gene g in samples S_a and S_b, we first note that each TSS-Seq tag is associated with a TSS locus where the first nucleotide of the tag is mapped. We denote $D_{a,g}$ and $D_{b,g}$ as the sets of TSS loci for gene g that have at least one observed TSS tag in S_a and S_b, respectively. Let us further denote $H_g = D_{a,g} \cup D_{b,g}$, and $|H_g| = N_g$. We can then construct a $2 \times N_g$ contingency table where each column of the table corresponds to a TSS locus that has at least one observed TSS tag in either S_a or S_b for gene g, and the rows correspond to the two samples. Each entry of the contingency table is the number of tags observed at the corresponding locus and sample, where the joint distribution of raw TSS tag counts at the N_g loci in H_g for each sample can be modeled as a multinomial distribution (Feller, 1968). The question of whether the TSS tags mapped to gene g are distributed differently in S_a and S_b can then be answered by testing the null hypothesis that tags mapped to g in S_a and S_b are distributed according to the same multinomial distribution.

Chi-square test for homogeneity is often applied to contingency tables to determine whether there exist differences in sample distribution. However, applying chi-square test in a global sense to all entries of a contingency table will only be able to determine whether the distribution of TSS tags at the possible TSS loci is different in the two samples being compared. Therefore, if the null hypothesis is rejected by the chi-square test, the only conclusion that can be reached is that the TSS tag distributions of gene g in the two samples most likely did not come from the same distribution. The lack of local information on the proportionality differences is insufficient for the purpose of detecting possible alternative promoter usage, where indications of the locations and direction of proportion changes are required to describe the pattern of change. To facilitate an automated and systematic approach to the detection of different patterns of TSS distribution changes, we first define two patterns of distribution changes that are of interest to us when comparing a gene's TSS distributions in two different samples: Class 1, where the peaks of TSS tag distribution are at the same position for both samples, but the peak contains a larger proportion of tags in S_b than in S_a; and Class 2, where the peaks of TSS tag distribution are located at different positions in S_a and S_b. With this categorization, Class 1 should contain those genes that are differentially expressed but using the same promoter in the two samples being compared, and Class 2 will contain those genes that may be candidates for differential promoter usage. In order to classify the TSS tag distribution change of gene g between S_a and S_b to Class 1 or Class 2, our proposed classifiers will partition the observed TSS loci in H_g into smaller subsets, where each subset can be individually tested so that local changes in proportion and the directions of change may be identified.

Let us denote the rows of a $2 \times N_g$ contingency table as $\rho_{a,g}$ and $\rho_{b,g}$, where each row is distributed as a multinomial distribution given its row sum. To perform local analysis on TSS loci in H_g, the $2 \times N_g$ contingency table can be partitioned into subtables, each of which is tested against the null hypothesis that both rows of the subtable come from the same distribution. To determine which of the loci in H_g have changes in proportion, we would like to have the number of subtables to be in the same range as the number of loci in H_g, where one subtable would correspond to one locus in H_g. For the proposed method, we partition the $2 \times N_g$ contingency table into N_g-1 subtables of size 2×2, where the left column contains one column of count data from the original contingency table, and the right column contains the sum of the observed tags in loci that have not been partitioned yet. The corresponding loci for the subtables and the pattern of acceptance and rejection of the null hypotheses are compared to our definitions for Class 1 and 2 distribution change patterns, and then combined to give a final significance score for the classification. The partitioning of a generic $2 \times C$ contingency table is illustrated in Fig. 1. The p-values of the tests on these subtables are combined to give an aggregate p-value for the significance of the classification.

For 2×2 contingency tables, Fisher's exact test is often used to determine significance of difference in proportion. It should be noted that the directions of proportion change at the TSS loci are important indicators for the classification of TSS distribution change. Here we propose a pattern for TSS distribution changes that is indicative of differential expression and alternative promoter usage, i.e., there is a statistically significant increase in proportion at the peak locus with respect to the total tags for g in S_b when compared to the proportion at the same locus in S_a, accompanied by statistically significant decreases in proportions at all other loci. Whether the distribution change indicates differential expression or alternative promoter usage will depend on the relative locations of the peaks in the two samples. To capture the different directions of proportion changes, one-sided Fisher's exact tests are used in the proposed classifier instead of the two-sided Fisher's exact test. The one-sided Fisher's exact test can be used to compare the ratio between the two entries in the top row, against the ratio between the two entries in the bottom row, with the null hypothesis that the two rows have the same ratio, or an odds ratio of 1. By appropriately setting

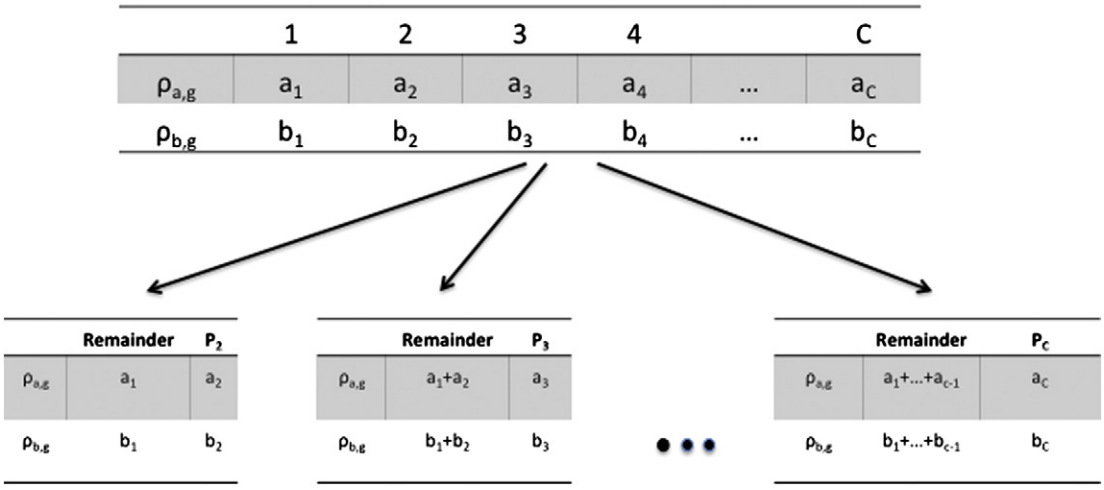

Fig. 1. Method for partition a 2 × N contingency table into (N-1) independent 2 × 2 contingency tables.

the one-sided Fisher's exact test to test for the alternative hypothesis of having odds ratio of greater than 1 or less than 1, depending on whether the subtable corresponds to the peak locus in S_b or to one of the remaining loci, we can perform local comparisons of the TSS distribution differences between S_a and S_b for the alternative hypothesis which is the pattern model for our distribution change classes. We now define the pattern model for Class 1 and Class 2 changes based on the directions of proportion changes in the subtables:

Class 1. For gene *g*, the locus with the largest TSS tag count in S_a and the locus with the largest TSS tag count in S_b, denoted as $P_{a,max}$ and $P_{b,max}$, respectively, are at the same position, i.e., $P_{a,max} = P_{b,max} = P_{max}$. Proportion of locus tag count to the total tag count observed for *g* in sample S_b is higher than that in sample S_a for the subtable corresponding to P_{max}. Proportion of locus tag count at each of the remaining subtables is lower in S_b than that in S_a.

Class 2. For gene *g*, the locus with the largest TSS tag count in S_a and the locus with the largest TSS tag count in S_b, denoted as $P_{a,max}$ and $P_{b,max}$, respectively, are at different positions, i.e., $P_{a,max} \neq P_{b,max}$. For gene *g*, proportion of locus tag count to the total tag count in sample S_b is higher than that in S_a for the subtable corresponding to $P_{b,max}$. Proportion of locus tag count at each of the remaining subtables, including $P_{a,max}$, is lower in S_b than that in S_a.

For each of the N_g-1 subtables, one-sided Fisher's exact test can be used to test either the "less than" or "greater than" alternative hypothesis, depending on the overall pattern and the location of the corresponding locus for the subtable. To obtain a single significance score that will indicate how closely the overall distribution change of *g* follows the pattern for the proposed Class 1 or Class 2 distribution changes, these p-values need to be aggregated into a single score. We first note that our method of partitioning the 2 × N_g contingency table in fact produces independent contingency tables of size 2 × 2 (Lancaster, 1949). This means that we are able to use approaches such as Fisher's method or Stouffer's method for combining p-values (Fisher, 1925; Stouffer et al., 1949), which require that the test statistics be independent to combine the p-values. However, typically, H_g will contain mostly loci with small tag counts. If all subtables are treated equally, p-values for loci with very small tag counts may be treated with the same importance and overwhelm contributions from p-values of the few loci with large tag counts. In order to properly consider the contribution of each locus based on its observed tag counts in S_a and S_b, we can apply Lipták's (1958) method, which is Stouffer's method with weights, to combine the N_g-1 p-values. Lipták's method for combining N p-values is given as follows:

$$Pg = 1 - \Phi\left(\frac{\sum_{i=1}^{N} w_i Z_i}{\sqrt{\sum_{i=1}^{N} w_i^2}}\right)$$

where $Z_i = \Phi^{-1}(1 - p_i)$ and p_i is the p-value of the *i*th subtable, w_i is the weight of the *i*th subtable and is equal to the maximum of the observed TSS tag counts at locus *i* in samples S_a and S_b, and Φ is the cumulative distribution function of a standard normal distribution.

Typically, transcription start sites associated with a promoter are distributed in the vicinity of the promoter. Transcripts that began their transcription at these sites would still be utilizing the same promoter. In other words, a gene that is found to have a statistically significant change in peak location between samples may not necessarily be a good candidate for a gene with alternative promoter usage if the change in distance is small. The classifier for Class 2 genes should therefore allow the user to specify a minimum distance, denoted as d_{min}, where only changes in peak locations with distances greater than or equal to d_{min} will be considered as utilizing different promoters. Furthermore, it also makes sense to merge the tag counts of all TSS that utilize the same promoter when testing for change in proportion between samples. We will use a windowed approach by assuming that all loci within a given window of size d_{win} centered on the peak locus are using the same promoter as the peak locus does. Since the window is applied to peaks in both samples, we have the constraint that $d_{min} > d_{win}$. To take into account of the merged subtables, the partitions will be made according to Fig. 2.

It should also be noted that some genes might have very low tag counts at all observed loci, and the classification for these genes will be highly affected by the amount of noise present. Even if there were indeed some biological signals present, the small amount of observed tags will typically have no biological significance. Therefore, we will also require that there are at least t_{min} observed tags at each of the peaks. Putting everything together, the distribution change classifiers for Class 1 and Class 2 are now given by the following pseudo code:

(1) Find $P_{a,max}$, the locus with the largest number of tag counts in S_a. Construct the first 2 × 2 subtable according to Fig. 2. Apply one-sided Fisher's exact test to test whether the proportion is increased from S_a to S_b at $P_{a,max}$.

(2) For the remaining loci in H_g, construct 2 × 2 subtables according to Fig. 2, in decreasing order of average observed tag counts.

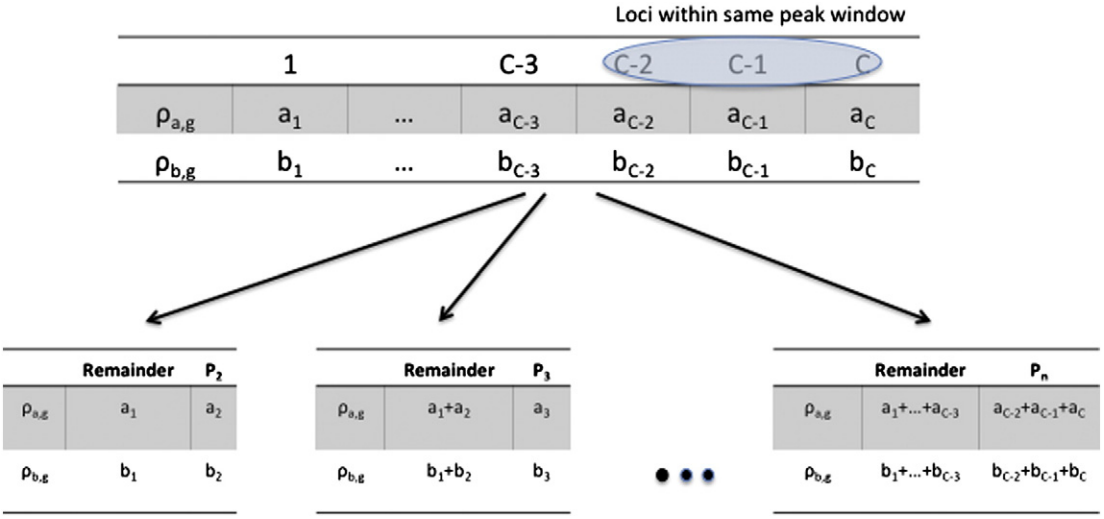

Fig. 2. Partitioning of 2 × N contingency table in to independent 2 × 2 contingency tables with loci within window merged into a single peak.

Apply one-sided Fisher's exact test for each subtable to test whether the proportion is decreased from S_a to S_b.

(3) Combine the resulting p-values with Lipták's method, using the average observed tag counts in S_a and S_b for a given locus as the weight of that locus.

(4) For genes with different peak locus in S_a and S_b, find $P_{a,max}$, the locus with the largest number of tag counts in S_a, and $P_{b,max}$, the locus with the largest number of tag counts in S_b. Construct the first 2 × 2 subtable according to Fig. 2 with sum of the tag counts at $P_{b,max}$ and loci within a window of d_{win} base pairs centered on $P_{b,max}$. Apply one-sided Fisher's exact test to test whether the proportion is increased from S_a to S_b at $P_{b,max}$.

(5) Construct 2 × 2 subtable with the sum of the tag counts at $P_{a,max}$ and loci within a window of size d_{win} base pairs centered on $P_{a,max}$. Apply one-sided Fisher's exact test to test whether the proportion is decreased from S_a to S_b at $P_{a,max}$.

(6) For the remaining loci in H_g, construct 2 × 2 subtables in decreasing order of average observed tag counts. Apply one-sided Fisher's exact test for each subtable to test whether the proportion is decreased from S_a to S_b.

(7) Combine the resulting p-values with Lipták's method, using the average observed tag counts in S_a and S_b for a given locus as the weight of that locus.

It should be noted that the proposed framework for classifying TSS distribution changes can also be easily modified to detect additional patterns of TSS distributions, such as genes that utilize one TSS region in one sample, but multiple TSS regions in the other. Detection of these patterns can be achieved by using Fisher's exact test with alternative hypotheses of increasing proportion for the subtables corresponding to the peaks in the second sample, and alternative hypotheses of decreasing proportion for the remaining subtables.

2.2. Extension for time-course analysis

As stated in the Introduction, an important motivation for the classifiers is for the understanding of how TSS usage in dendritic cells is affected by an immune response. The response of a gene to stimulation is a dynamic event that can be analyzed by studying the time-course data obtained from samples taken at various time points before and after the initial stimulation.

For a time-course experiment with N sampling points, let us assume that there is 1 sample taken before stimulation, and the remaining N-1 samples are taken after stimulation. To understand the time-course response of a gene's TSS distribution to stimulation, we can make multiple comparisons between samples taken after stimulation to the sample taken before stimulation to determine the difference between TSS distribution at each time point and that from the original, pre-stimulated sample. The classifiers designed to search for Class 1 and 2 patterns are applied to each of the N-1 comparisons. We can reduce the likelihood of finding genes with significant Class 1 or Class 2 TSS distribution change patterns due to noise by making sure that the pattern of change remains consistent throughout the N-1 comparisons. Similar to how the p-values of all the subtables in the classifiers are combined, Lipták's method is again used to combine p-values of the N-1 tests to give a single value representing the overall significance of a gene having Class 1 and Class 2 distribution changes over the duration of the experiment. For Class 1 patterns, we can simply combine the p-values from the N-1 comparisons using Lipták's method. However, for Class 2 patterns, we want to select genes that not only consistently show Class 2 change patterns, but also have consistent shifts to the same genomic region. To achieve this, we take the peak locus with the largest TSS tag count in the N-1 time points after stimulation as the center locus. If the peak locus of any of the remaining N-2 samples falls outside of the window of size d_{win} around the center locus, its p-value for the comparison with the sample before stimulation will be set to 1. This ensures that only those samples with shifts to the same general vicinity will be counted as evidence for the Class 2 distribution change pattern. The weights used in the Lipták's method for combining the N-1 comparisons will be the gene's total TSS tag counts in the sample taken after LPS stimulation. To ensure that the tag counts in each sample is comparable, we will normalize the tag counts into parts per million (PPM).

2.3. Within-sample variation

Here we present an approach to account for a possible effect of within-sample variation on the detection of genes with significant TSS distribution changes. In a comparison of two different samples, differences in expression found between the samples can either be attributed to biological differences between the samples, or to differences due to other factors such as natural variation within the same tissue or even sample preparation. In order to ensure that the genes found to be significant exhibit these differences in TSS distributions due to biological differences, we propose a simulation approach here to remove genes that are more prone to showing significant TSS distribution changes due to the within-sample variation.

Due to Poisson distribution's limitation in that its mean and variance are equal, it often cannot account for variations in expression that occur

in biological replicates. Instead, negative binomial distribution has been proposed to address this problem of over-dispersion (Anders and Huber, 2010; Robinson et al., 2010). Here, we will also use negative binomial distribution to generate simulated biological replicates for real data with a single sample. For each locus with observed TSS tags in the real dataset, we will generate a simulated tag count for the simulated dataset using a negative binomial distribution $NB(\mu, \mu + \mu^2/\delta)$, where μ is equal to the number of tags observed at the locus, and δ is the dispersion parameter. TSS distribution change pattern classifiers are then applied to a comparison of real and simulated dataset to find genes with significant distribution change with the specified patterns that may occur due to within-sample variation.

2.4. Material

To evaluate the performance and usefulness of the proposed pattern classifiers, we applied the classifiers to two sets of data. The first set of data was obtained from the repository at DBTSS TSS-Seq data for human adult tissues of brain, colon, heart, kidney, liver and lymph. A total of 30 pair-wise comparisons are made to identify genes with tissue-specific differential expressions and alternative TSS usages.

The second set of data is a time-course experiment on mouse dendritic cells to determine the changes in TSS landscape after an immune response. We used mouse dendritic cells extracted from bone-marrow cells with the presence of GM-CSF. These dendritic cells are stimulated with lipopolysaccharide (LPS) to elicit immune response. Samples are collected at 0 h, 0.5 h, 1 h, 2 h, 3 h, 4 h, 6 h, 8 h, 16 h, and 24 h from LPS stimulation, and TSS-Seq was used to map out the TSS distribution landscape for each of the 10 samples.

2.4.1. Human tissues

The six human adult tissues available from DBTSS have already been mapped to RefSeq hg19 reference genome (Pruitt et al., 2009). The mapped locations of the reads are compared to the genomic coordinates of RefSeq genes. We will assume that those reads whose 5′-ends were located within the second or the later exons are likely the results of re-capping of erroneously truncated transcripts (Suzuki et al., 2001). In order to minimize the number of incorrectly identified TSS, only those reads whose 5′-ends were located 10 kb upstream and within the first exon, and those that were mapped within the intronic regions were kept for subsequent analysis (Kimura et al., 2006).

Class 1 and Class 2 classifiers are applied to all possible pairs of human adult tissues. We used $d_{win} = 30$ and $d_{min} = 100$ base pairs to obtain gene lists for significant Class 1 and Class 2 genes. We also set $t_{min} = 10$ to filter out genes that were inactive or had very low expressions in one or both of the samples. Since the proposed classifiers are directional, different orderings of the same pair of tissues will give different results. Therefore, a total of 30 pairs are tested for changes in TSS distribution. The list of significant genes is filtered using the within-sample variation correction approach proposed in Section 2.3 using 50 simulated datasets with the dispersion parameter $\delta = 10$. A plot of the coefficient of variance (ratio between the mean and standard deviation) for mean values of 1 to 5000 can be found in Supplementary B. A gene is discarded for a tissue if any one of the 50 simulated comparisons for the tissue result in a smaller p-value for the corresponding pattern classifier. Furthermore, for the remaining genes in each adult tissue, only those genes that have false discovery rate (FDR) of less than 0.05 when compared to all remaining tissues using Class 1 and Class 2 classifiers are retained for subsequent analysis. In other words, we only analyze genes that have unique expression pattern or TSS usage in a specific tissue in this study.

2.4.2. Mouse dendritic cell

TSS-Seq reads from each of the 10 time samples were mapped to the RefSeq mm9 mouse reference genome (Pruitt et al., 2009) using the short read aligner, Bowtie (Langmead et al., 2009). Those reads that were not perfectly and uniquely mapped to the reference genome were removed from the datasets. Furthermore, to remove those reads that may be the results of re-capping, we again keep only those reads that are mapped to 10 kb upstream and within the first exon, and those that are mapped to the intronic regions. The mapping statistics for the 10 samples are summarized in Supplementary A.

To analyze the time-course dendritic cell dataset for TSS distribution changes after LPS stimulation, we use the proposed classifiers with the time-course extension to generate lists of genes with significant Class 1 and Class 2 changes. Parameters used for mouse dendritic cell analysis are the same as the ones used in the analysis for human tissues. To remove genes possibly having significant distribution changes due to within-sample variations, we generate simulated time-course datasets, each of which consists of the sample taken before LPS stimulation, and 9 simulated samples generated from it using negative binomial distribution with $\delta = 10$. Genes with consistent Class 1 and 2 pattern changes across the simulated time course dataset are found in a similar manner to the real dataset. A total of 50 such time-course simulations are performed, and genes with p-value for simulated data that is more significant than p-value for real data in 1 or more simulations are removed from the list of significant genes obtained from the real dataset. FDR is then applied to each list for multiple comparison correction.

3. Results/discussion

3.1. Human tissue

The lists of genes with unique Class 1 and Class 2 changes for the 6 human adult tissues were analyzed using DAVID (Huang et al., 2009a, 2009b) to find statistically overrepresented gene groups, gene ontology terms and tissues with similar expression profile. Using the proposed TSS distribution change classifiers, we generated two gene lists for each tissue: 1) genes that have Class 1 TSS distribution change pattern when compared to the same genes in all other tissues, and 2) genes that have Class 2 TSS distribution change pattern when compared to the same genes in all other tissues. In other words, List 1 contains genes that are more highly expressed or have tissue-specific expression, but show no signs of alternative TSS usage compared to the remaining tissues. On the other hand, List 2 contains genes that utilize unique TSS region in the tissue that was tested compared to all other tissues in the test. The two lists identify genes that are most unique for each tissue in terms of differential expression and TSS usage shift.

Tables 1 and 2 show the top categories in gene group (SP-PIR keywords), gene ontology (Biological Process) and tissue expression (Uniprot Tissue) found to have statistically significant Class 1 and Class 2 distribution change pattern in each tissue. For all three comparisons, we select the top 3 categories in terms of FDR. Comparing the number of genes found, we see that for all 6 tissues, there are more genes in Class 2 lists than in Class 1 lists. In particular, human adult brain tissue has the largest number of genes found to have both significant Class 1 and Class 2 changes, suggesting that genes in brain tissues have the most unique transcriptional landscape of all the tissues tested here. Indeed, for brain tissue Class 1 and Class 2 genes, the significant terms or groups in each of the 3 analyses show a very good match to known functions that are specific to brain. In Pardo et al. (2013), KCNIP4, PCDH9, CADM2, BAI3, NRG3, LSAMP, NRXN1, LRRTM4, and FGF14 were identified to be outliers, and have more than 100 alternative TSS clusters. All genes were found in the Class 2 gene list for brain tissue in our study except for LSAMP and FGF14, which are left out of brain's unique Class 2 list for having FDR > 0.05 in the comparison between brain and colon tissues. In Motojima and Goto (1989), transthyretin (TTR) is also found to be using different promoter regions in liver and brain. In our comparisons, if we use p-value cutoffs instead, TTR is also shown to have significant Class 2 distribution change pattern between liver and brain, with a p-value of 1.84E − 3.

Table 1
Overrepresentation analysis for uniquely expressed genes in human adult tissues with no alternative TSS usage.

Brain (109)	Colon (24)	Heart (24)	Kidney (31)	Liver (36)	Lymph (38)
Gene group (SP-PIR keywords)					
Synapse	Acetylation	Electron transport	Acetylation	Acetylation	Ribosomal protein
p = 3.2E−8	p = 2.2E−3	p = 6.3E−3	p = 1.3E−5	p = 3.2E−4	p = 3.1E−21
Postsynaptic cell membrane	Ribonucleoprotein	Acetylation	Transit peptide	Oxidoreductase	Ribosome
p = 1.9E−7	p = 4.3E−3	p = 3.0E−2	p = 3.7E−2	p = 4.1E−4	p = 5.9E−21
Cell junction	cGMP	Methylation	Mitochondrion	Oxidative phosphorylation	Ribonucleoprotein
p = 3.8E−7	p = 2.5E−2	p = 3.4E−2	p = 3.9E−2	p = 1.7E−3	p = 1.3E−18
Gene ontology (biological process)					
Trans. nerve impulse	Steroid metabolic process	Response to inorganic substance	Organelle localization	Oxidation reduction	Translational elongation
p = 8.1E−6	p = 3.5E−2	p = 2.8E−3	p = 1.2E−2	p = 1.4E−3	p = 2.1E−24
Synaptic trans.	Response to drug	Protein complex assembly	Mitochondrion localization	Generation of precursor metabolites and energy	Translation
p = 1.2E−5	p = 4.0E−2	p = 4.8E−3	p = 2.4E−2	p = 3.2E−3	p = 1.1E−19
Ion transport	N/A	Protein complex biogenesis	Negative regulation of cell proliferation	Positive regulation of catalytic activity	N/A
p = 1.9E−4		p = 4.8E−3	p = 2.8E−2	p = 3.3E−3	
Tissue expression (Uniprot Tissue)					
Brain	Colon adenocarcinoma	Heart	Cajal–Retzius cell	Liver	Colon adenocarcinoma
p = 2.1E−7	9.1E−2	3.3E−2	3.1E−2	p = 1.9E−4	p = 1.1E−2
Fetal brain	N/A	Liver	Colon carcinoma	Urinary bladder	Lymph
p = 1.4E−3		3.7E−2	3.0E−2	p = 5.8E−3	p = 4.3E−2
Cerebellum	N/A	Lung	Muscle	Fetal brain cortex	Blood
p = 1.4E−3		8.9E−2	3.3E−2	p = 7.0E−3	p = 4.5E−2

Each column lists the top 3 terms from gene group, gene ontology and tissue expression analyses, and their corresponding p-value below for each of the human adult tissues. Numbers in parenthesis beside the tissue names are the number of genes in that tissue that are found to have Class 1 TSS distribution change when compared to other tissues.

For gene ontology analysis, we can see that several tissues are found to be significant in terms that match the tissues' known functions. For example, for brain tissue, both Class 1 and 2 genes are significantly represented in terms associated with transmission of nerve impulses. In heart, terms related to cardiac muscles and muscles in general are overrepresented with Class 1 genes, and catabolic processes are found to be significant for both liver and kidney Class 1 genes. On the other hand, while some of the most significant terms may not be tissue-specific, they may also be important to the tissue's functions. For example, in kidney, positive regulation of I-κB kinase/NF-κB cascade is significantly represented with Class 2 genes, and it is known that NF-κB is an important regulator of kidney inflammation (Sanz et al., 2010).

In terms of tissue expression profile, several tissues also show statistically significant matches with Uniprot Tissue data. In our comparison, Class 1 lists appear to do a better job than Class 2 lists at identifying the tissues. For Class 1 lists, all tissues except for kidney and lymph are correctly identified. Note that since there are only 6 tissues used in this analysis, it is possible that an untested tissue may appear as a significant

Table 2
Overrepresentation analysis for genes in human adult tissues with unique alternative TSS usage.

Brain (302)	Colon (221)	Heart (47)	Kidney (57)	Liver (82)	Lymph (55)
Gene group (SP-PIR keywords)					
Acetylation	Phosphoprotein	Acetylated amino end	Acetylation	Ribosomal protein	Mitochondrion
p = 1.8E−11	p = 3.4E−13	p = 1.7E−3	p = 2.7E−3	p = 1.0E−6	p = 1.1E−5
Ribosome	Acetylation	Acetylation	Mitochondrion	Ribonucleoprotein	Respiratory chain
p = 7.3E−10	p = 3.2E−7	p = 2.0E−3	p = 6.1E−3	p = 2.1E−6	p = 4.0E−5
Protein biosynthesis	Cytoplasm	Copper	Lysosome	Acetylation	Transit peptide
p = 3.1E−9	p = 1.3E−6	p = 7.9E−3	p = 7.2E−3	p = 3.1E−6	p = 2.8E−4
Gene Ontology (Biological Process)					
Translational elongation	Striated muscle tissue development	Generation of precursor metabolites and energy	Positive regulation of I-kB/NF-kB cascade	Translational elongation	Generation of precursor metabolites and energy
p = 1.4E−9	p = 6.1E−4	p = 1.0E−3	p = 2.5E−2	p = 3.3E−8	p = 2.5E−4
Translation	Muscle cell differentiation	Muscle system process	Regulation of I-kB/ NF-kB cascade	Translation	Electron transport chain
p = 1.9E−7	p = 6.7E−4	p = 8.5E−3	p = 2.9E−2	p = 3.5E−5	p = 4.3E−3
Negative regulation of protein ubiquitination	Phosphorylation	Muscle thin filament assembly	Cell−cell adhesion	Regulation of myeloid leukocyte differentiation	Mitochondrial electron transport, NADH to ubiquinone
p = 2.4E−4	p = 6.7E−4	p = 1.3E−2	p = 3.2E−2	p = 1.7E−2	p = 6.5e−3
Tissue Expression (Uniprot Tissue)					
Brain	Colon Carcinoma	Skeletal muscle	Liver	Colon	Skin
p = 3.9E−7	p = 3.5E−5	p = 4.2E−4	p = 1.3E−6	p = 1.2E−3	p = 3.0E−3
Amygdala	Epithelium	Urine	Bone marrow	Liver	Ovary
p = 2.9E−5	p = 5.6E−5	p = 2.0E−3	p = 3.7E−3	p = 4.2E−3	p = 2.6E−2
Cerebellum	Brain	Platelet	Plasma	Umbilical cord blood	Lung
p = 3.4E−4	p = 8.0E−5	p = 2.1E−2	p = 3.7E−3	p = 6.9E−3	p = 4.4E−2

Each column lists the top 3 terms from gene group, gene ontology and tissue expression analyses, and their corresponding p-value below for each of the human adult tissues. Numbers in parenthesis beside the tissue names are the number of genes in that tissue that are found to have Class 2 TSS distribution change when compared to other tissues.

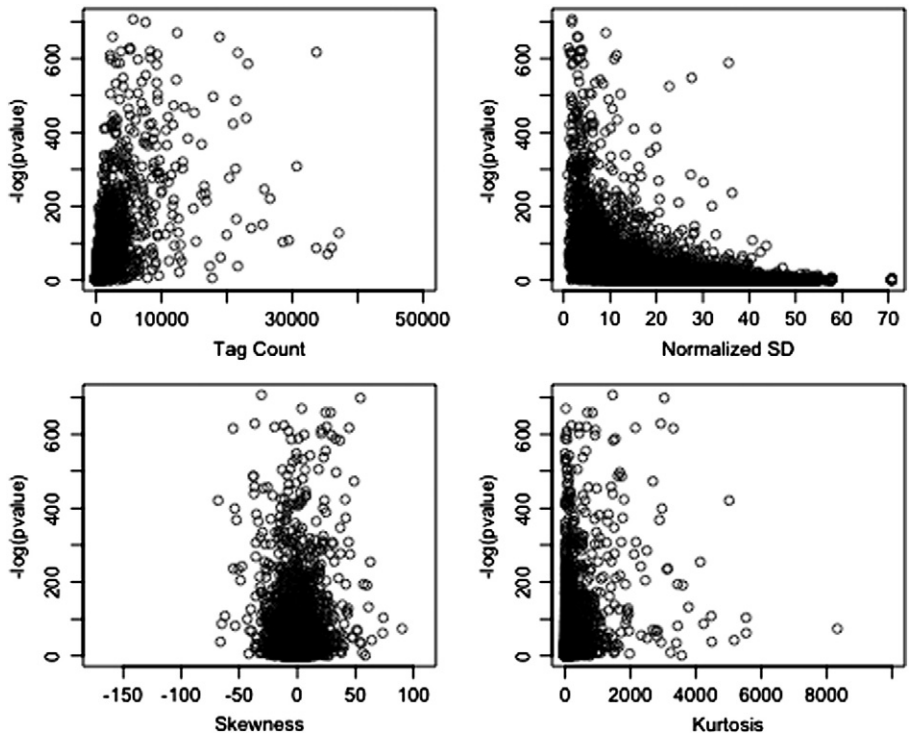

Fig. 3. Plots of total tag count, normalized standard deviation, skewness, and kurtosis vs. p-value for simulated Class 1 time-course dataset.

match to a tested tissue's TSS profile if their TSS profiles are very similar. The two tissues could be discriminated if the TSS profile of the untested tissue could be obtained. It should be expected that Class 2 genes do not perform as well in the identification of a tissue as Class 1 genes do. Due to the specification of its TSS distribution change pattern, Class 2 can include not only genes that have increased expression, but also genes that have a relatively similar level of expression, as long as they exhibit signs of significant alternative TSS usage. These genes that have alternative TSS usage but similar expression levels across different tissues may become confounding factors in identifying tissues through tissue-specific expression.

By comparing the gene groups, gene ontology terms and tissue expression profiles between the 6 human adult tissues and between List 1 and List 2 genes, it is clear that even when using such restrictive

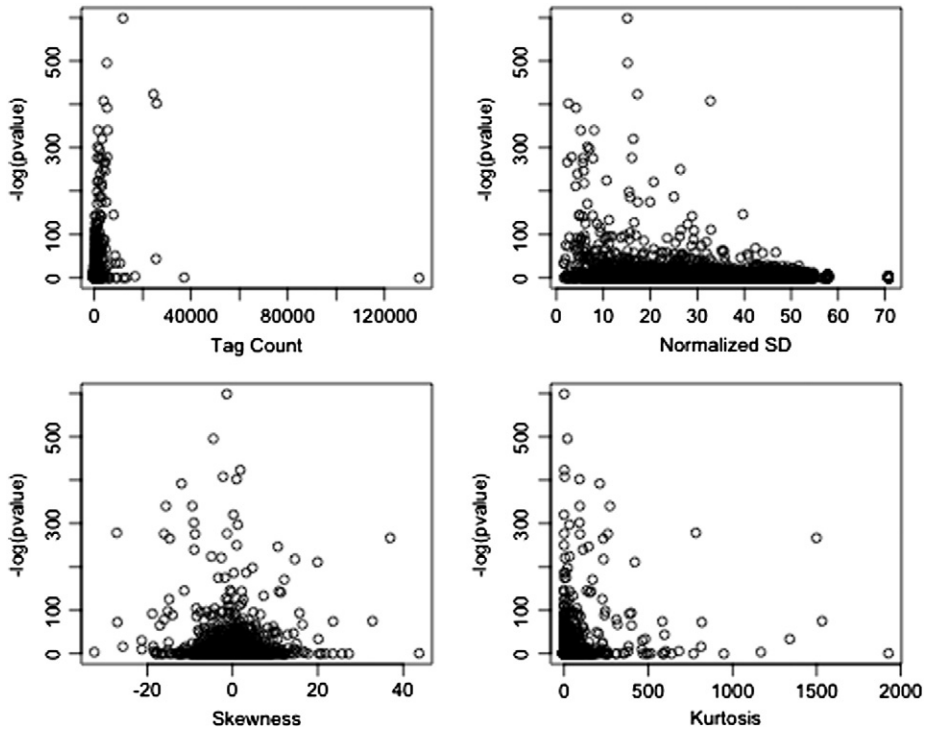

Fig. 4. Plots of total tag count, normalized standard deviation, skewness, and kurtosis vs. p-value for simulated Class 2 time-course dataset.

Table 3
Immune-process related Class 1 and Class 2 genes for time-course mouse dendritic cell experiment.

Class 1	MyD88-dependent	MyD88-independent	NF-κB	Chemokines	
	Gadd45g Irg1 Nfkbiz Socs1	Ifit1 Isg15 Usp18	Tnf Zc3h12c	Cxcl1 Cxcl2 Cxcl3 Cxcl16 Ccl3	
Class 2	Both MyD88-dependent and TRIF-dependent				
	Il6	Il27	Stat5a	Ikzf1	Trafd1

Class 1 genes are divided into 4 categories: genes involved in MyD88-dependent pathway, genes involved in TRIF-dependent pathway, NF-κB genes and chemokines. Class 2 genes are divided into two categories: genes involved in both MyD88- and TRIF-dependent pathways, and other.

requirements such as selecting unique Class 1 and Class 2 genes in each tissue, clear differences can been seen between the two classes and between the tissues. Additional similarities and differences between genes can be found by relaxing the uniqueness requirement to find tissues whose genes may share the same alternative TSS usage, and shed light on the common pathways or mechanisms that maybe used in those tissues.

3.2. Mouse dendritic cells

3.2.1. Simulated data and within-sample variation

To have a better understanding of what kinds of TSS tag distributions are more likely to have lower p-values from the simulation analyses, we plotted p-values of TSS distributions from the simulation study to the moments of those distributions. Fig. 3 shows the plots of the minimum Class 1 p-value from the 50 simulated iterations for the time-course analysis versus total tag count, standard deviation, skewness, and kurtosis of the TSS tag distributions from the real samples taken before LPS stimulation. In these plots, only genes with p-values less than 1 are plotted. Out of a total of 14,316 genes, 11,161 genes have a minimum p-value less than 1. Of the 4 plots, only normalized SD show a significant trend for p-value. From the plot we can see that low p-values are associated with lower standard deviation, i.e., the less spread out its TSS distribution is, the more likely it is for a gene to show spurious Class 1 distribution change pattern due to within-sample variation.

Fig. 4 shows the same plots for p-values from Class 2 pattern classifications for 4948 genes with a minimum p-value less than 1. From the figures, we can see that total tag count does not show a strong correlation with p-value. While there appears to be some trend of decreasing p-value with increasing total gene tag count, there are also a not insignificant number of cases where genes with large tag counts have very high p-values for Class 2 distribution change pattern. For standard deviation, we can see that most of the genes with extremely small p-values

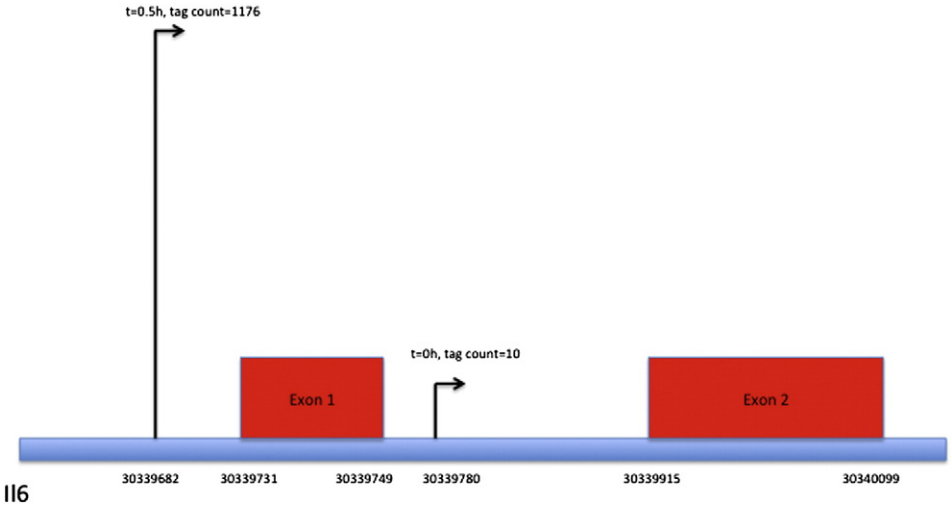

Fig. 5. Positions of Il6 TSS peak in samples taken at t = 0 h and t = 0.5 h with respect to exon locations.

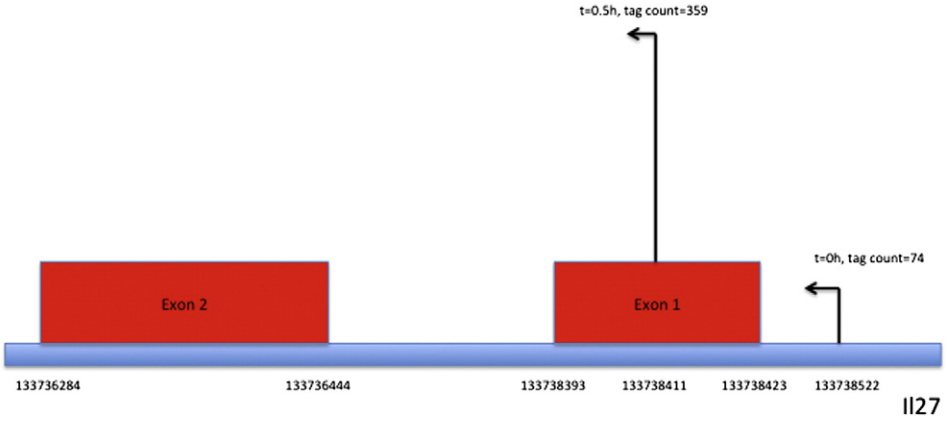

Fig. 6. Positions of Il27 TSS peak in samples taken at t = 0 h and t = 0.5 h with respect to exon locations.

Table 4
Gene expression peak time for immune-process related Class 1 and Class 2 genes.

	0 h	0.5 h	1 h	2 h	3 h	4 h	6 h	8 h
1		Nfkbiz	Ccl9 Cxcl1 Cxcl2 Tnf Zc3h12c Irg1	Ccl3 Cxcl3	Ifit1 Irf1	Socs1	Isg15 Usp18	Gadd45g
2		Ikzf1		Il6 Il27	Sgpl1 Stat5a Trafd1			

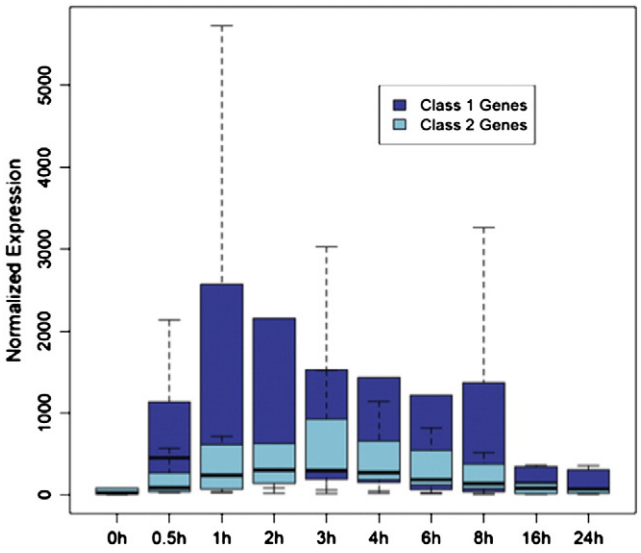

Fig. 7. Overlaid comparison of box-plots for expressions of Class 1 and Class 2 immune-related genes, chemokines and NF-κB genes.

can be found in the region with normalized SD less than 30. Skewness and kurtosis appear to show a better correlation with the p-value, where low p-value genes are found to have smaller absolute skewness and also smaller or even negative kurtosis. The relationship between kurtosis and p-value can be understood intuitively in that small kurtosis typically indicates a less peaked or even multimodal distribution. If a gene has a less peaked or multimodal distribution of TSS tags across all of its loci, then a high within-sample variation at the individual loci can easily lead to a significant p-value for Class 2 distribution change pattern in the simulated data or in biological replicates.

3.2.2. Analysis of TSS distribution change

For real data, 1671 genes were found to have a p-value less than 0.05 for Class 1 distribution change pattern, and 921 remain after removing genes found to have simulated p-value more significant than that of the real dataset. For Class 2, 288 genes were found to have a p-value less than 0.05, and 242 remain after filtering. After correcting for multiple comparisons using FDR, we are left with 671 genes for Class 1 and 215 genes for Class 2.

Genes that are related to the TLR-signaling pathways, chemokines, and NF-κB genes from Class 1 and 2 lists are shown in Table 3. Several of the genes classified as Class 2 genes, Il6, Il27, Stat5a, Ikzf1, and Trafd1 are involved in MyD88- and TRIF-dependent signaling (Molle et al., 2007). On the other hand, Class 1 genes such as Gadd45g, Irg1, Nfkbiz, and Socs1 (p-value cutoff) are involved in the MyD88-dependent pathway, and Class 1 genes such as Isg15, Usp18, and Ifit1 are involved in the TRIF-dependent pathway. Comparing the classification of the genes and their dependency on MyD88-dependent and TRIF-dependent pathways, it appears that there may be some connection between how a gene is controlled by the signal pathways and its promoter usage, where different signaling pathways may be utilizing different promoters for the transcription of the same gene. We see that the genes that are known to be associated with only one of the signaling pathways are also classified as Class 1 genes, which indicates that these genes are found to be using the same TSS before and after LPS stimulation. As for Il6, Il27, Ikzf1, and Trafd1, genes that are known to be associated with both signaling pathways are classified as Class 2 genes, where the classifier has detected a statistically significant shift in peaks between samples taken before and after LPS stimulation. In Fig. 5 we plotted the peaks in samples taken at $t = 0$ h and $t = 0.5$ h against the exon positions for Il6. Before the LPS stimulation, the peak TSS locus is located in the intronic region between the first and second exons, but the peak TSS locus moved upstream of the first exon to include the first exon in the transcript in the sample taken at $t = 0.5$ h after the LPS stimulation. In Fig. 6, a similar shift in TSS peak locus can be observed for Il27. In the sample taken at $t = 0.5$ h, we can see that the peak locus has shifted from upstream of the first exon to within the first exon. The shift of TSS peak locus across significant genomic structures after LPS stimulation, coupled with the known fact that both of these genes are associated with two different signaling pathways indicate that there may indeed be some biological significance behind the change in TSS distribution.

An estimation of a gene's expression at a given time period can be obtained by adding together all the observed TSS tags for that gene in the corresponding sample. After normalizing all samples into ppm, we can plot the time-course expression of the genes at the 10 sample points. In Table 4, the Class 1 and Class 2 genes are rearranged by the time when each gene's expression has reached. While there are relatively fewer number of Class 2 genes, Table 4 still shows that a larger portion of Class 1 genes reach their maximum expression early in the reaction to LPS stimulation (between 0.5 h and 1 h), and the Class 2 genes reach their maximum expression at a later period (between 2 h and 3 h). The boxplot of the time-course expression for the two classes in Fig. 7 shows the same trend. Both Table 4 and Fig. 7 suggest that there may be some correlation between changes in TSS usage and the duration and delay of the response to LPS stimulation.

Table 5
Top 10 significant gene ontology terms for Class 1 and Class 2 genes found in time-course mouse dendritic cell experiment.

Class 1 GO term	Class 1 p-value	Class 2 GO term	Class 2 p-value
Immune response	1.5E−5	Positive regulation of T cell differentiation	3.3E−5
Generation of precursor metabolites and energy	7.7E−5	Positive regulation of lymphocyte differentiation	4.5E−5
Response to bacterium	6.3E−4	Positive regulation of immune system process	1.6E−4
Cell proliferation	8.3E−4	Regulation of T cell differentiation	2.4E−4
Positive regulation of apoptosis	8.6E−4	Regulation of lymphocyte differentiation	5.9E−4
Positive regulation of programmed cell death	9.1E−4	Positive regulation of T cell activation	1.1E−3
Cell homeostasis	9.3E−4	Positive regulation of leukocyte regulation	1.1E−3
Induction of programmed cell death	9.5E−4	Leukocyte activation	1.1E−3
Induction of apoptosis	9.5E−4	Positive regulation of cell activation	1.2E−3
Positive regulation of cell death	9.7E−4	Positive regulation of alpha-beta T cell activation	1.4E−3

3.2.3. Gene ontology and gene group analysis

To determine whether there are any significant functional differences between the Class 1 and Class 2 genes, we obtained the significant gene ontology (GO) terms for both Class 1 and Class 2 lists by uploading the lists to DAVID and obtained the significantly over-represented terms in the GOTERM_BP_FAT annotation category which includes GO terms related to biological processes. Table 5 lists the top 10 most significant GO terms for Class 1 and Class 2 genes.

Since both lists contain genes that are activated after the LPS stimulation, it is not surprising that the top significant GO terms for both gene lists are related to immune processes. In Table 5, we can see that the top significant terms for Class 1 genes are the top level GO categories for immune processes, whereas for Class 2 genes, the top significant terms are more focused in terms of dealing with the activation and differentiation of leukocytes, particularly T cells, which are an important component of adaptive immune response. For Class 1 gene list, the most significant term related to lymphocyte activation does not pass the 0.05 p-value cutoff, and has no terms related to lymphocyte differentiation at all. While there are Class 1 immune genes whose functions involve activation and differentiation of leukocytes, statistically, various aspects of immune process are more evenly represented so that no specific function stand out in terms of statistical significance. On the other hand, while Class 2 list contains only a few genes with immune process related GO terms, the majority of them appears to be involved in the activation and differentiation of leukocytes which leads to the higher significance of specific immune process terms.

In addition to gene ontology, we can also see significant differences in gene groups associated with Class 1 and Class 2 genes. Table 6 lists the significant gene groups from Swiss-PROT keywords for the two classes. We can see from the table that the alternative splicing gene group is significantly overrepresented by genes from Class 2, which indicates that many of the genes in Class 2 list not only show significant shift in their TSS usage location after LPS stimulation, but they are also known to produce multiple isoforms through the alternative splicing mechanism. It should also be noted that while not listed as statistically over-represented in the table, the zinc-finger gene group also includes Ikzf1 and Trafd1, found in Table 3 as immune-process related genes. The promoter of Ikzf1, or Ikaros family zinc finger protein 1, is shown to bind to the IRF gene family (Fang et al., 2012), which is involved in both the MyD88- and TRIF-dependent portions of the TLR4 signaling pathway. Trafd1, or TRAF-type zinc finger domain-containing protein 1, is a negative regulator that attenuates the MyD88-dependent NF-κB activation pathway and suppresses TRIF-mediated NF-κB activation (Mashima et al., 2005).

Table 7 lists the significant KEGG pathways for Class 1 and 2 genes. Interestingly, genes in JAK–STAT signaling pathway are found to be significantly overrepresented in Class 2 genes. JAK–STAT signaling pathway has been shown to interact with both the MyD88-dependent and TRIF-dependent parts of the TLR4 pathway. Recent studies have shown that MyD88 mutation can enhance JAK–STAT3 signaling (Poulain et al., 2013), and signaling through TRIF and IFNR activates a JAK–STAT pathway to induce expression of surface molecules required for the interaction with T cells (Brieger et al., 2013). Findings from gene ontology, gene group, and KEGG pathway analyses show that there is evidence of some genes in the TLR signaling pathways undergoing alternative TSS usage after LPS stimulation, and that this mechanism is found predominantly in genes involved in both sub-pathways, which may indicate a functional significance. While analysis through only TSS-Seq data is not conclusive evidence for such a mechanism, it does provide good candidates for subsequent investigations into whether the two possible signaling pathways utilize alternative promoter regions to induce the production of different functioning isoforms.

Table 6
Top 7 significant gene group terms for Class 1 and Class 2 genes found in time-course mouse dendritic cell experiment.

Class 1 gene group	Class 1 p-value	Class 2 gene group	Class 2 p-value
Acetylation	8.0E−15	Alternative splicing	3.3E−4
Isopeptide bond	3.3E−4	Cytoplasm	1.5E−3
Hydrogen ion transport	1.6E−3	Phosphoprotein	3.5E−3
Cytoplasm	1.7E−3	Heterodimer	9.6E−3
Phosphoprotein	4.1E−3	Activator	1.8E−2
Chemotaxis	5.3E−3	Lysosome	2.1E−2
Nucleus	7.8E−3	Cytokine receptor	2.2E−2

Table 7
Top 3 significant KEGG pathways for Class 1 and Class 2 genes found in time-course mouse dendritic cell experiment.

Class 1 KEGG pathway	Class 1 p-value	Class 2KEGG pathway	Class 2 p-value
Oxidative phosphorylation	1.9E−3	JAK–STAT signaling pathway	2.4E−3
Amino sugar and nucleotide sugar metabolism	3.0E−3	Lysosome	3.1E−3
Lysosome	2.4E−2	Cytokine–cytokine receptor interaction	7.6E−2
Huntington's disease	3.9E−2		
Amyotrophic lateral sclerosis	4.3E−2		

4. Conclusion

In this work we have proposed a classifier for the automatic discovery of genes that exhibit differential TSS usage under different conditions or in different tissues. By reconfiguring the classifier, we can discover genes with specific patterns of distribution change that are of interest to the kind of biological phenomenon that the researchers are interested in. In this work, we showed that the proposed classifiers could pick out genes that are found in different gene groups, gene ontology terms and tissue expression profiles with statistical significance using TSS-Seq data for human adult tissues. More importantly, however, the classifiers can be used to find candidate genes that might be using alternative promoters as a mechanism to produce different isoforms. As shown by the analysis of LPS-stimulated mouse dendritic cells, we can see that an interesting difference can be observed between genes that are classified as Class 1 and Class 2 genes. The immune-related genes being classified as Class 1 and Class 2 genes are particularly interesting when compared to their involvement in the TLR4 signaling pathway. The classifier results have shown that those immune-related genes that are involved in only one of MyD88-depedent or TRIF-dependent pathways tend to be classified as Class 1 genes, whereas those immune-related genes that are classified as Class 2 genes are typically involved in both pathways. Furthermore, by comparing the genomic locations of the peak loci before and after stimulation for these Class 2 genes, it can be seen that some of these shifts occur across important genomic structures. Although whether alternative promoters have been used in these cases has to be confirmed through upstream transcription factor and ChIP-Seq analysis, classification using only TSS data has provided interesting targets for subsequent analysis.

The approach we have taken here can also be extended to test for more complicated forms of TSS distribution changes and construct different TSS usage profiles. So far, we have only tested for the model that only a single peak will have an increase in proportion when comparing two samples. If the peaks are at the same locus, the gene is classified as a Class 1 gene. If the peaks are at different loci separated by more than a predetermined distance, the gene is then classified as a Class 2 gene. In this model, any other side peaks with an increase in proportion are treated as noise and will increase the p-value of the gene. However, there may be cases where the increase in proportion at the side peaks may have a biological meaning, such as when multiple promoters are used at the same time to transcribe different isoforms. These situations can be easily handled by changing the direction of the

one-sided Fisher's exact test to take into account the side peaks as an important biological phenomenon. Furthermore, the proposed classifier is currently designed to work with experiments with one sample. While datasets with biological replicates can be easily handled by using the replicates to estimate the negative binomial parameters used in creating the simulated datasets, we would like to modify our approach to integrate the biological replicates in the statistical tests for distribution change as a future update for this method.

Conflict of interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2013.12.038.

References

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biol. 11, R106.

Bajic, V.B., Seah, S.H., Chong, A., Zhang, G., Koh, J.L., Brusic, V., 2002. Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics 18 (1), 198–199.

Balwierz, P.J., Carninci, P., van Nimwegen, E., et al., 2009. Methods for analysing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol. 10, R79.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799–816.

Brieger, A., Rink, L., Haase, H., 2013. Differential regulation of TLR-dependent MyD88 and TRIF signaling pathways by free zinc ions. J. Immunol. http://dx.doi.org/10.4049/jimmunol.1301261.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., et al., 2005. The transcriptional landscape of the mammalian genome. Science 309, 1559–1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Kazuro, S., et al., 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. 38 (6), 626–635.

Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C., Huang, T.H., 2008. The functional consequences of alternative promoter use in mammalian genomes. Trends Genet. 24, 167–177.

Down, T.A., Hubbard, T.J., 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res. 12, 458–461.

Fang, C.M., Roy, S., Pitha, P.M., et al., 2012. Unique contribution of IRF-5-Ikaros axis to the B-cell IgG2a response. Genes Immun. 13 (5), 421–430.

Feller, W., 1968. 3rd ed. An Introduction to Probability Theory and Its Applications, vol. I. John Wiley & Sons, Inc., New York.

Fisher, R.A., 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

Gustincich, S., Sandelin, A., Carninci, P., et al., 2006. The complexity of the mammalian transcriptome. J. Physiol. 575 (2), 321–332.

Hatchwell, E., Greally, J.M., 2007. The potential role of epigenomic dysregulation in complex human disease. Trends Genet. 23, 588–595.

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009a. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nat. Protoc. 4 (1), 44–57.

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009b. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37 (1), 1–13.

Jordan, I.K., Rogozin, I.B., Glazko, G.V., Koonin, E., 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 19, 68–72.

Kawai, T., et al., 2001. Lipopolysaccharide stimulates the MyD88-independent pathway and results in activation of IFN-regulatory factor 3 and the expression of a subset of lipopolysaccharide-inducible genes. J. Immunol. 167, 5887–5894.

Kawaji, H., Frith, M.C., Sandelin, A., et al., 2006. Dynamic usage of transcription start sites within core promoters. Genome Biol. 7, R118.

Kimura, K., Wakamatsu, A., Suzuki, Y., et al., 2006. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Genome Res. 16, 55–65.

Knudsen, S., 1999. Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics 15 (5), 356–361.

Lancaster, H.O., 1949. The derivation and partition of chi-square in certain discrete distributions. Biometrika 36, 117–129.

Landry, J.R., Mager, D.L., Wilhelm, B.T., 2003. Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet. 19, 640–648.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R20.

Lipták, T., 1958. On the combination of independent tests. Magyar Tud. Akad. Mat. Kutato Int. Közl. 3, 171–196.

Lu, J., Luo, L., 2008. Prediction for human transcription start site using diversity measure with quadratic discriminant. Bioinformatics 2 (7), 316–321.

Lu, Y.C., Yeh, W.C., Ohashi, P.S., 2008. LPS/TLR4 signal transduction pathway. Cytokine 42 (2), 145–151.

Mashima, R., Saeki, K., Aki, D., Yoshimura, A., et al., 2005. FLN29, a novel interferon- and LPS-inducible gene acting as a negative regulator of toll-like receptor signalling. J. Biol. Chem. 280 (50), 41289–41297.

Molle, C., Nguyen, M., Goriely, S., et al., 2007. IL-27 synthesis induced by TLR ligation critically depends on IFN regulatory factor 3. J. Immunol. 178 (12), 7607–7615.

Motojima, K., Goto, S., 1989. Brain-specific expression of transthyretin mRNA as revealed by cDNA cloning from brain. FEBS Lett. 258 (1), 103–105.

Ni, T., Corcoran, D.L., Rach, E.A., Song, S., Spana, E.P., et al., 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat. Methods 7, 521–557.

Okazaki, Y., et al., 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 457, 1038–1042.

Ota, T., et al., 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat. Genet. 36, 40–45.

Pardo, L.M., Rizzu, P., Francescatto, M., Carninci, P., Heutink, P., et al., 2013. Regional differences in gene expression and promoter usage in aged human brains. Neurobiol. Aging 34, 1825–1836.

Poulain, S., Roumier, C., Leleu, X., et al., 2013. MYD88 L265P mutation in Waldenstrom's macroglogulinemia. Blood. http://dx.doi.org/10.1182/blood-2012-06-436329.

Pruitt, K.D., Tatusova, T., Klimke, W., Maglott, D.R., 2009. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 37, D32–D36.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Rubner, Y., Tomasi, C., Guibas, L.J., 1998. A metric for distributions with applications to image databases. Sixth International Conference on Computer Vision,1998, pp. 59–66.

Sanz, A.B., Sanchez-Nino, M.D., Ramos, A.M., et al., 2010. NF-kB in renal inflammation. J. Am. Soc. Nephrol. 21 (8), 1254–1262.

Shen, H., Tesar, B.M., Walker, W.E., Goldstein, D.R., 2008. Dual signaling of MyD88 and TRIF is critical for maximal TLR4-induced dendritic cell maturation. J. Immunol. 181 (3), 1849–1858.

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., Williams Jr., R.M., 1949. The American soldier. Adjustment during Army Life, vol. 1. Princeton University Press, Princeton.

Suzuki, Y., et al., 2001. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep. 2, 388–393.

Takeda, K., Akira, S., 2004. TLR signaling pathways. Semin. Immunol. 16, 3–9.

The ENCODE Project Consortium, 2012. An integrated encyclopaedia of DNA elements in the human genome. Nature 489, 57–74.

Tsuchihara, K., Suzuki, Y., Wakaguri, H., et al., 2009. Massive transcriptional start site analysis of human genes in hypoxia cells. Nucleic Acids Res. 37, 2249–2263.

Yamashita, R., Sugano, S., Nakai, K., Suzuki, Y., et al., 2011. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome Res. 21 (5), 775–786.

Yamashita, R., Sugano, S., Suzuki, Y., Nakai, K., 2012. DBTSS: DataBase of Transcriptional Start Sites progress report in 2012. Nucleic Acids Res. 40 (D1), D150–D154.

Zhang, M.Q., 1998. Identification of human gene core promoters in silico. Genome Res. 8, 319-316.

Zhao, X., Valen, E., Parker, B.J., Sandelin, A., 2011. Systematic clustering of transcription start site landscapes. PLoS One 6 (8), e23409.